Are On-Line Personae Really Unlinkable?

What anybody can see....

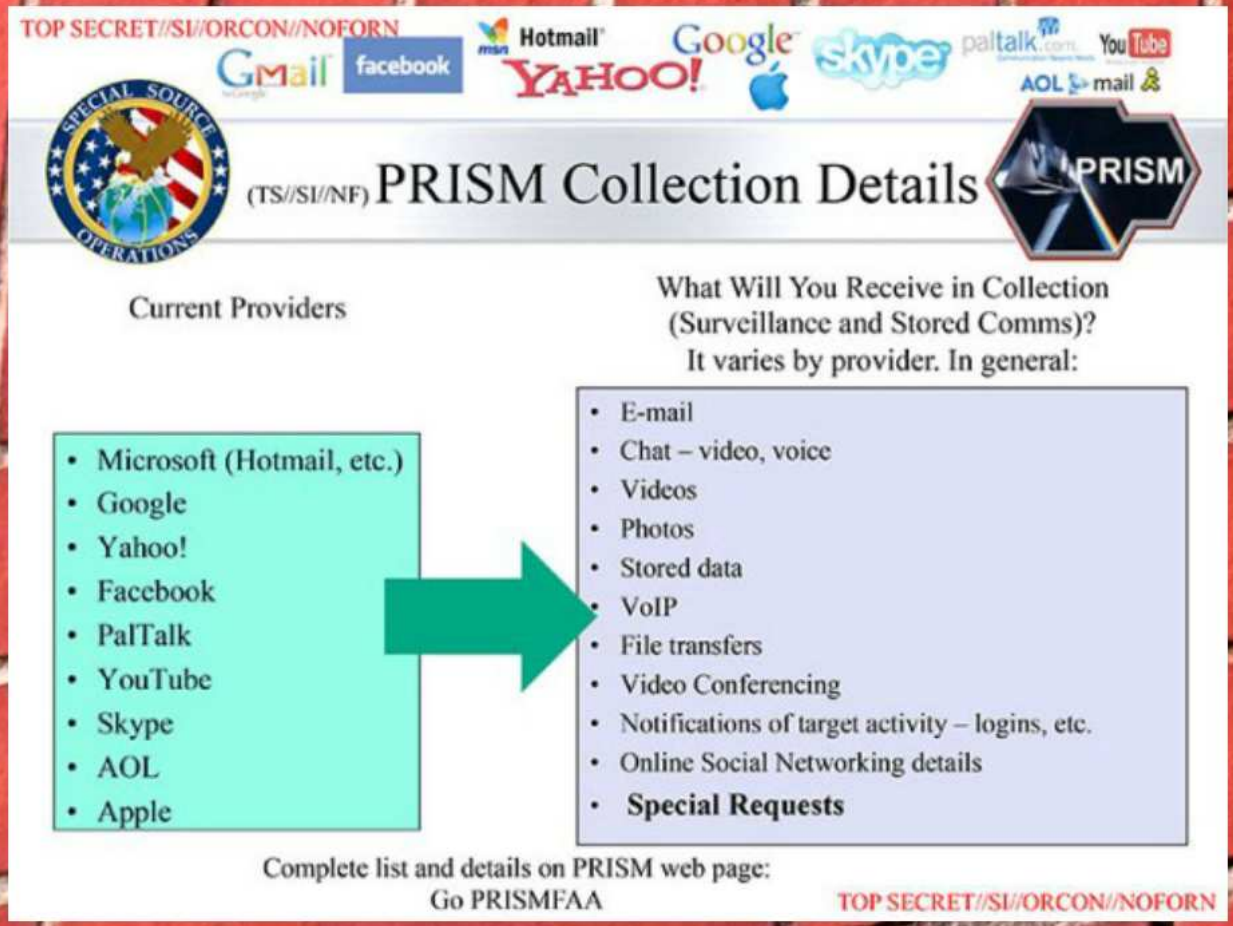
?

What the US government sees....

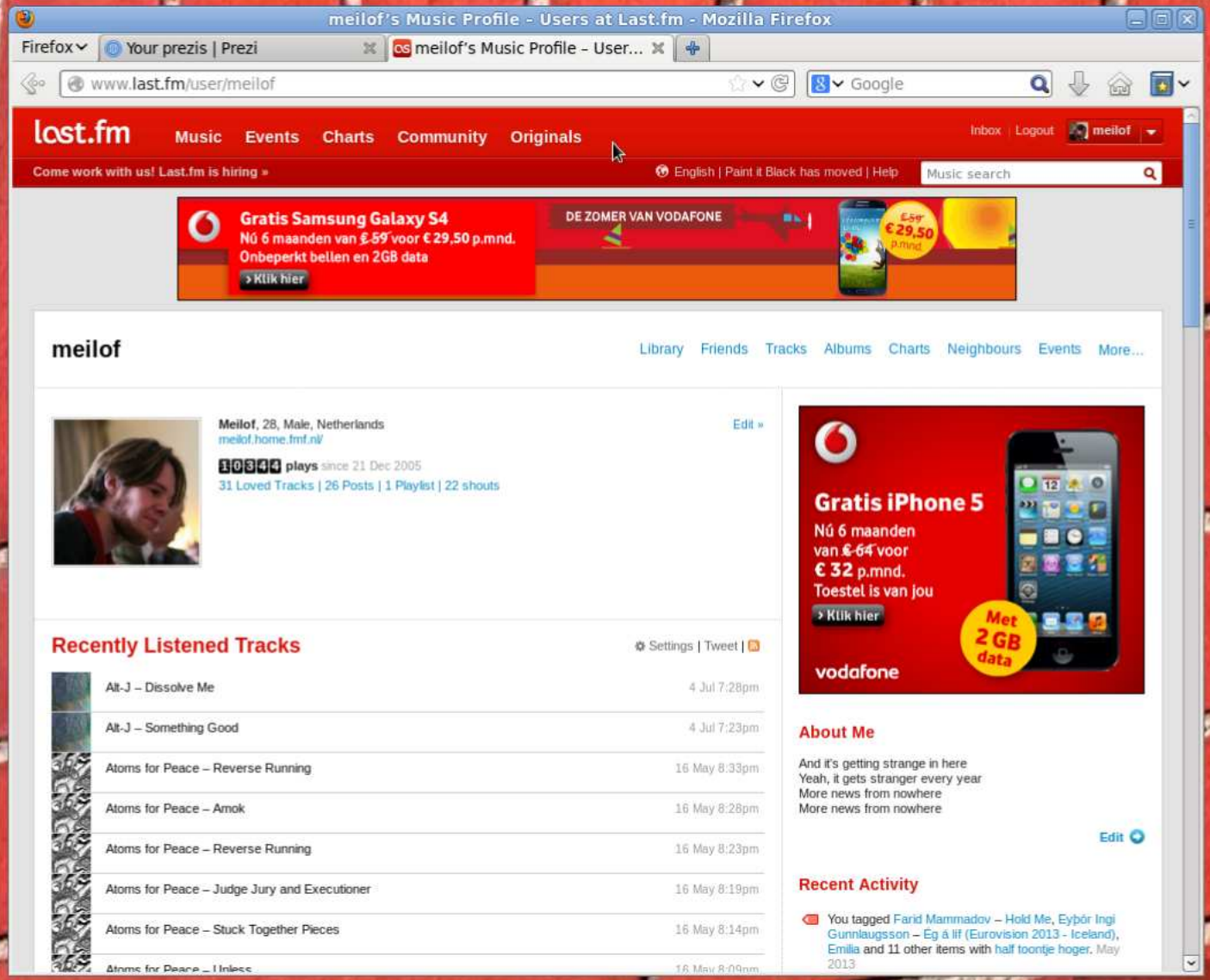
Record linkage:

common attributes, rare values

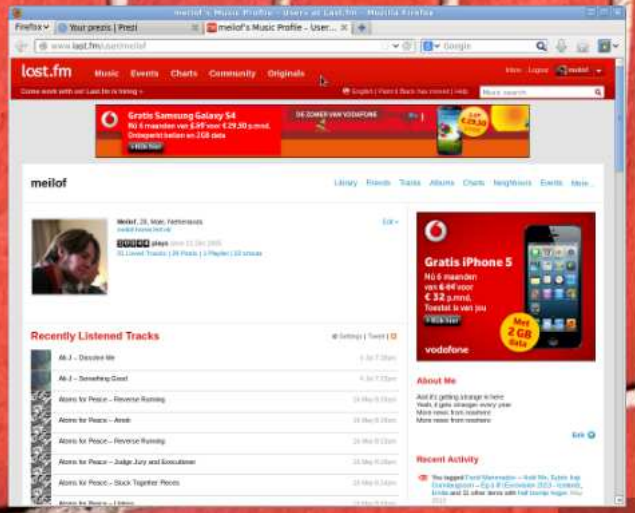
What the US government sees....



What anybody can see....

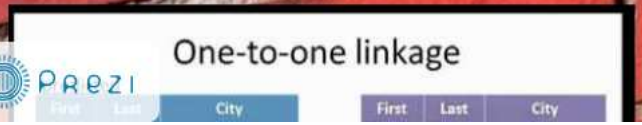


What anybody can see....



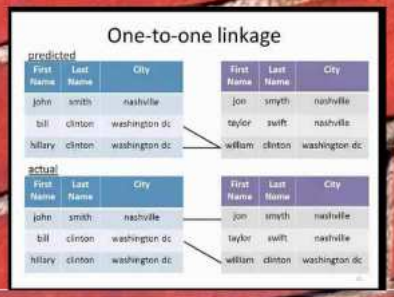
?

Record linkage:



What anybody can see....

?



Record linkage:

common attributes, rare values

What anybody can see....

?

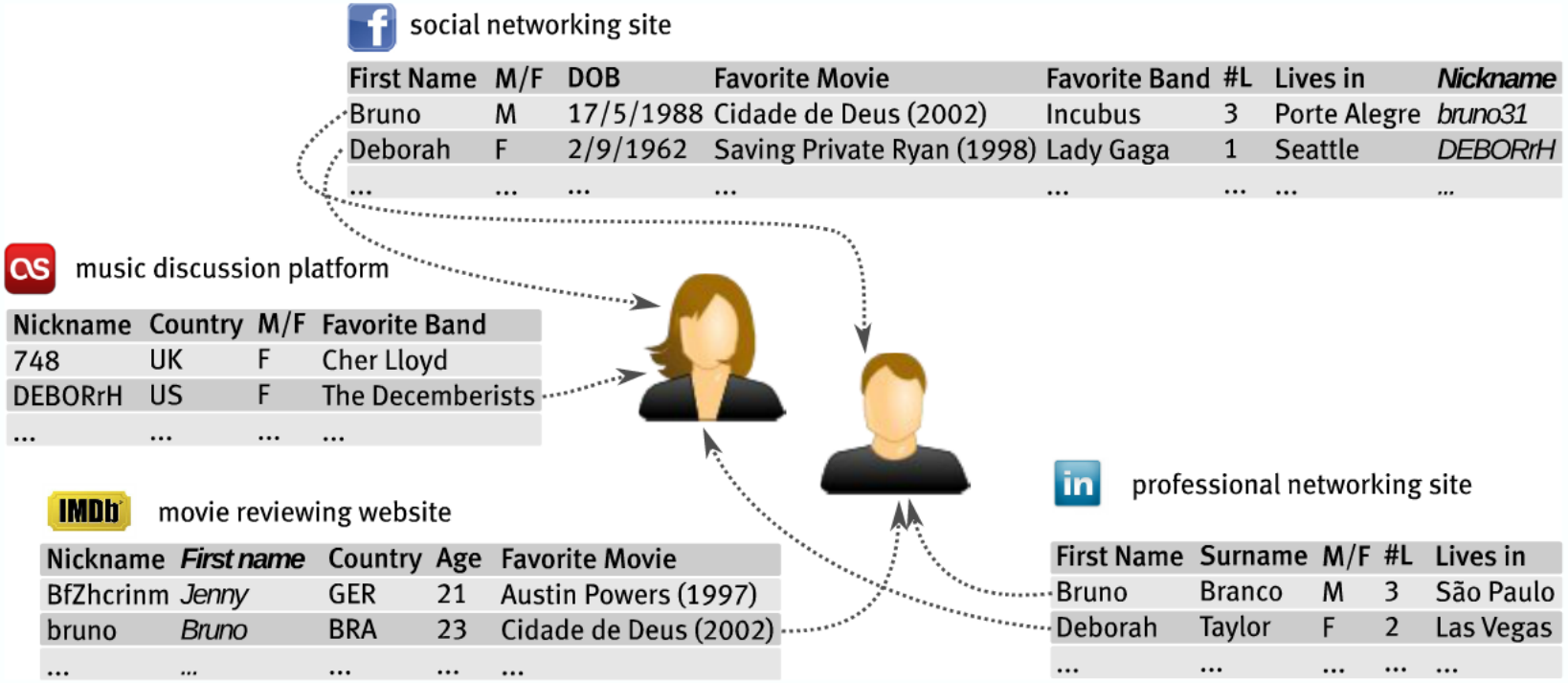
What the US government sees....

Record linkage:

common attributes, rare values

How big is the on-line privacy problem?

multiple databases



common attributes & values

List of most popular given names - Wikipedia, the free encyclopedia - Mozilla Firefox

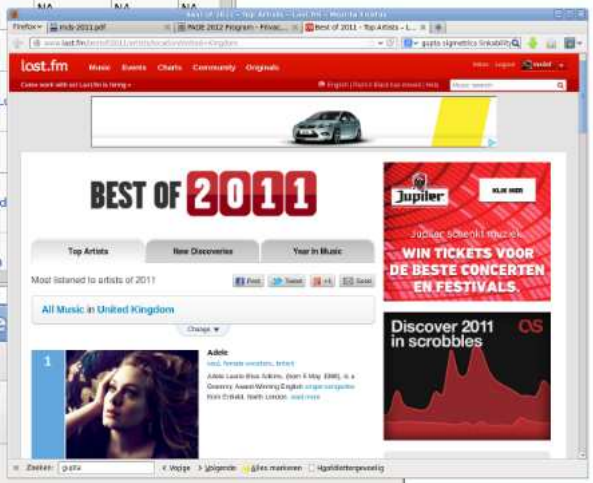
Firefox | Are On-Line Personae Really... | W List of most popular given n...

en.wikipedia.org/wiki/List_of_most_popular_given_names#North_and_South_America

North and South America [edit source | edit beta]

Male names [edit source | edit beta]

Region (year)	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No.
Buenos Aires, Argentina (2012)[100]	Juan	Benjamin	Santiago	Thiago	Lucas	Joaquín	Santino	Lautaro	Ian	Mateo
Aruba (2005)[101]	Daniel	Dylan/Dyllan	Kevin/Keven	NA	NA	NA	NA	NA	NA	NA
Brazil (2012)[102]	Miguel	Arthur	Davi	Gabriel	Lucas	Matheus	Pedro			
Canada (2012, unofficial list)[103]	William	Jacob	Liam	Nathan	Noah	Ethan	Lucas/L			
Canada, Alberta (2012)[104]	Liam	Ethan	Jacob	Logan	Mason	Benjamin	Lucas			
Canada, British Columbia (2012)[105]	Ethan	Liam	Lucas	Mason	Logan	Noah	Alexand			
Canada, Manitoba (2012)	Liam	Mason	Carter	Noah	Logan	Lucas	William			



*train.csv (/scratch/experiments/files/output_generator) - ge

File Edit View Search Tools Documents Help

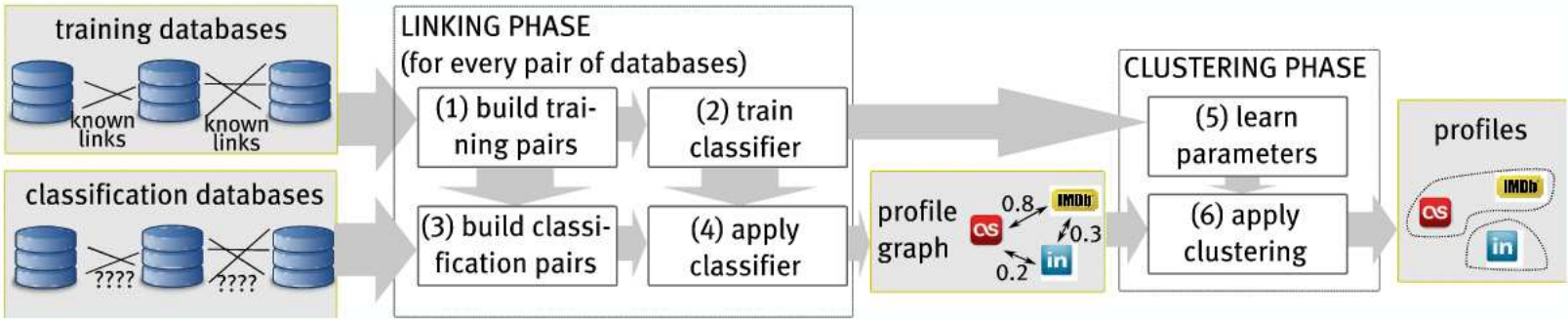
Open Save Undo

*train.csv

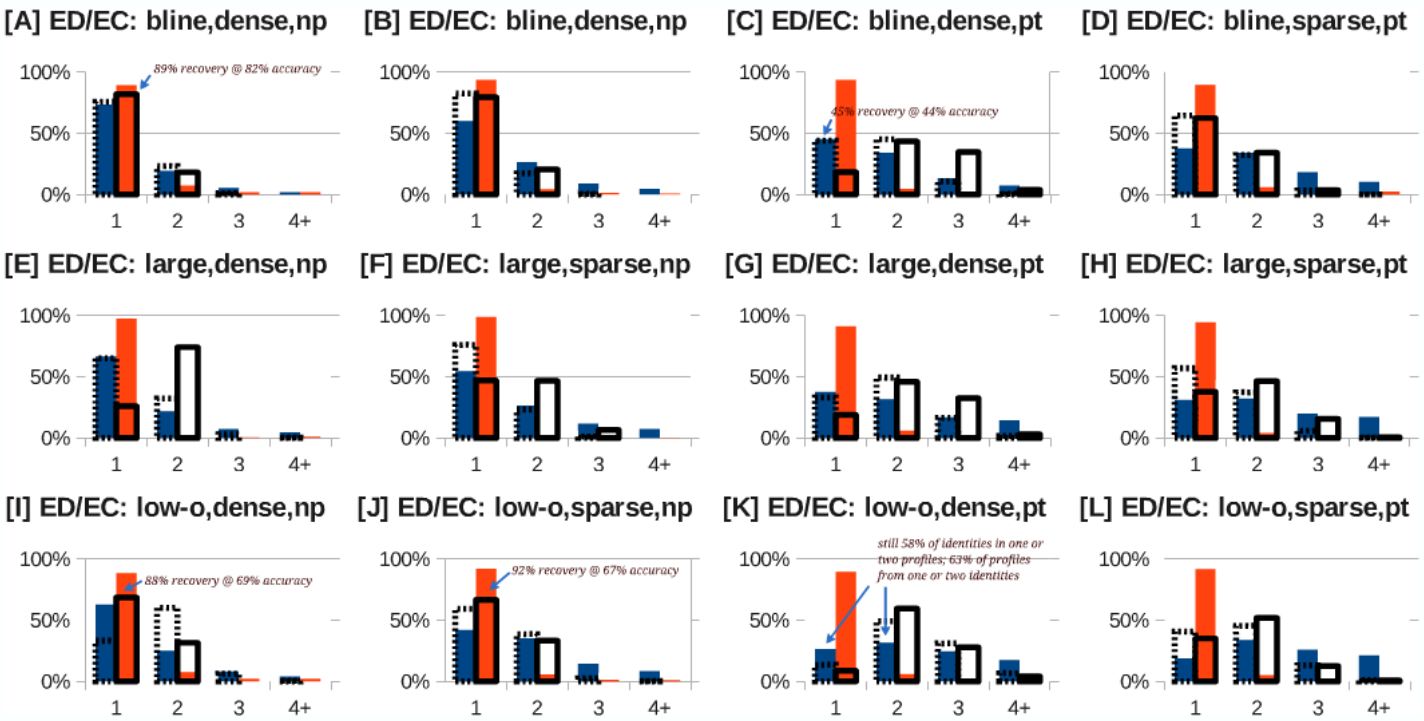
```
"RUTH","RUTH","RUTH","female","female","female","2","2","The Shawshank Redemption (1994)","Star Wars:
Episode V - The Empire Strikes Back (1980)","Bon Iver","Bon Iver","2","1","Colorado Springs","Colorado
Springs","us","us","us","us","RUTH","RUTHMImZaR ","RUTHMImZaR ","22","22"
"Seougia","Seougia","Seougia","female","female","female","21","21","The Avengers (2012)","The Avengers
(2012)","Britney Spears","Britney
Spears","4","4","London","London","uk","uk","uk","uk","yY","yY","yY","6","6"
"Rafaela","Rafaela","Rafaela","female","female","female","3","3","Orfeu Negro (1959)","Orfeu Negro
(1959)","Avril Lavigne","Radiohead","2","0","Porto
Alegre","Londrina","brazil","brazil","brazil","brazil","McR","RafdHla","McR","39","39"
```

Plain Text | Tab Width: 8 | Ln 1, Col 1 | INS

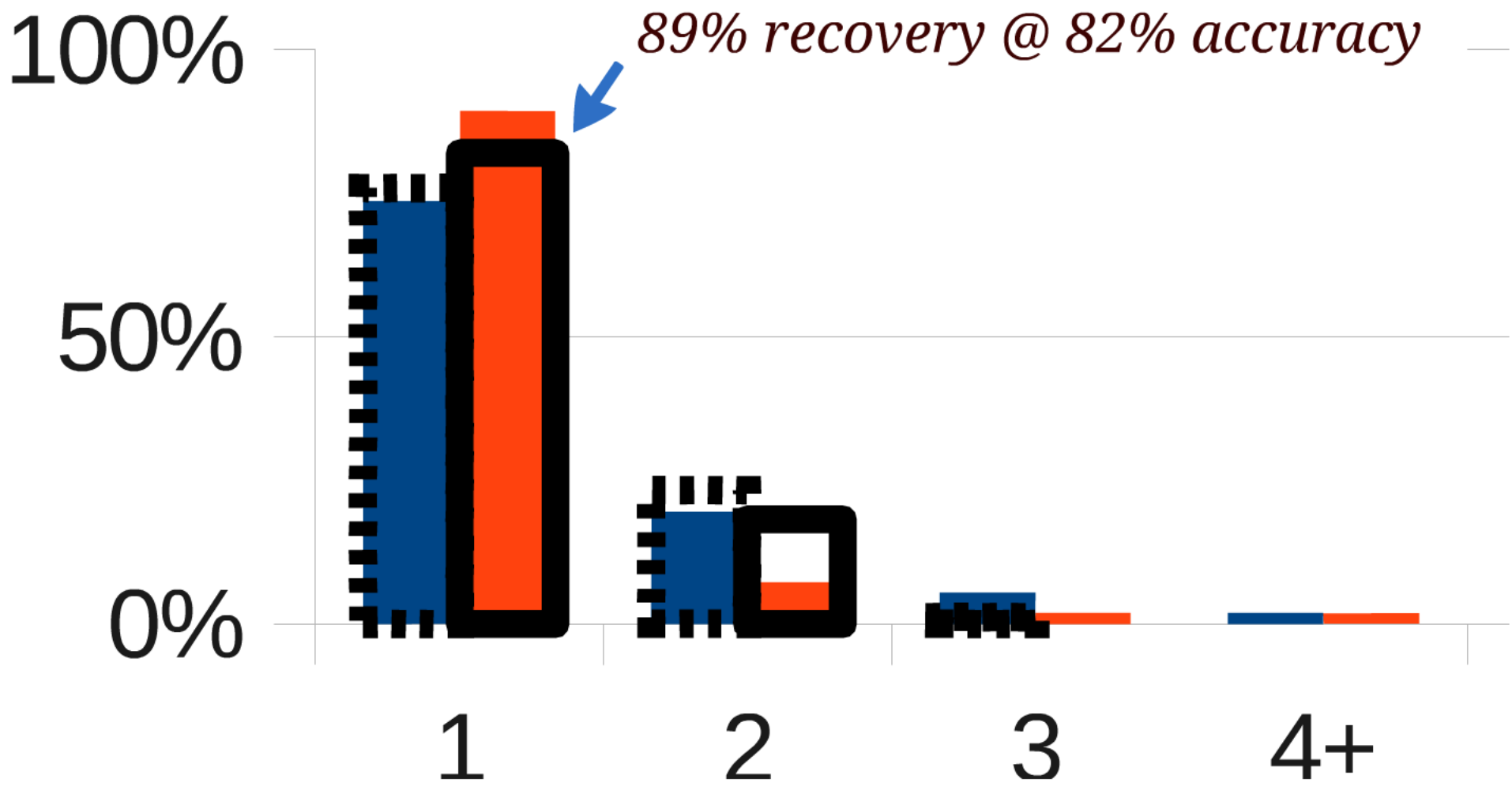
experimental numbers



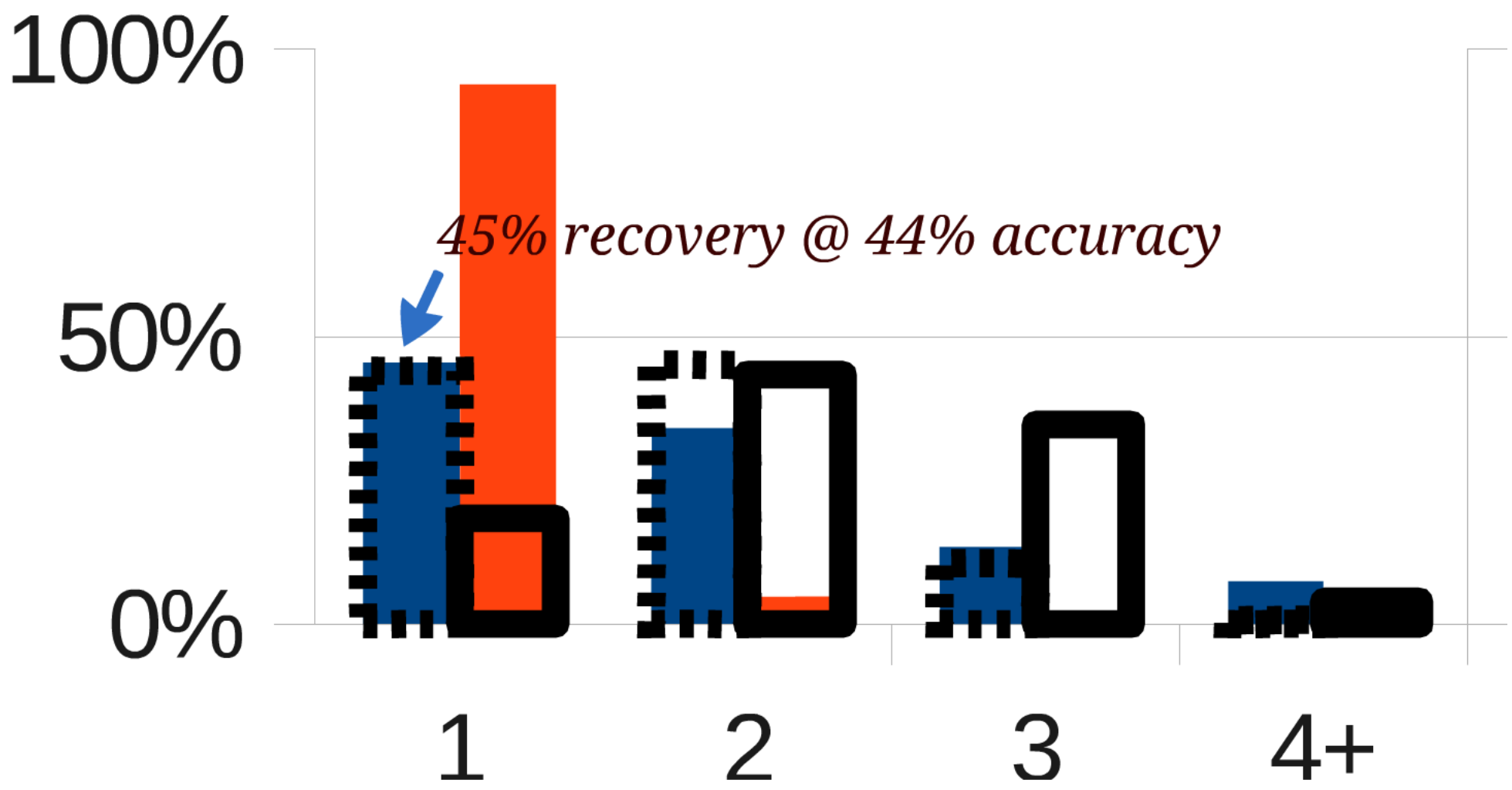
experimental numbers



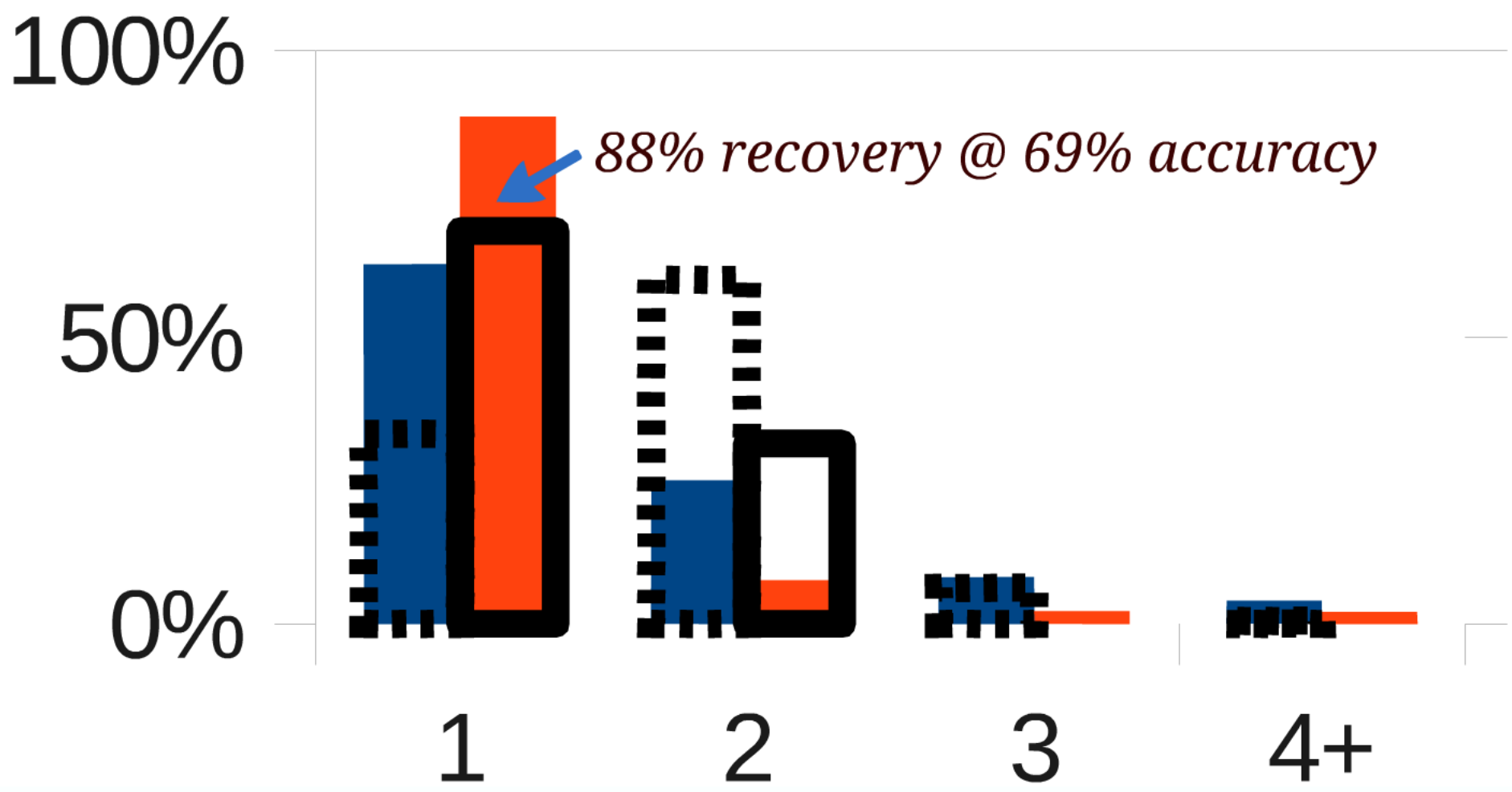
[A] ED/EC: bline,dense,np



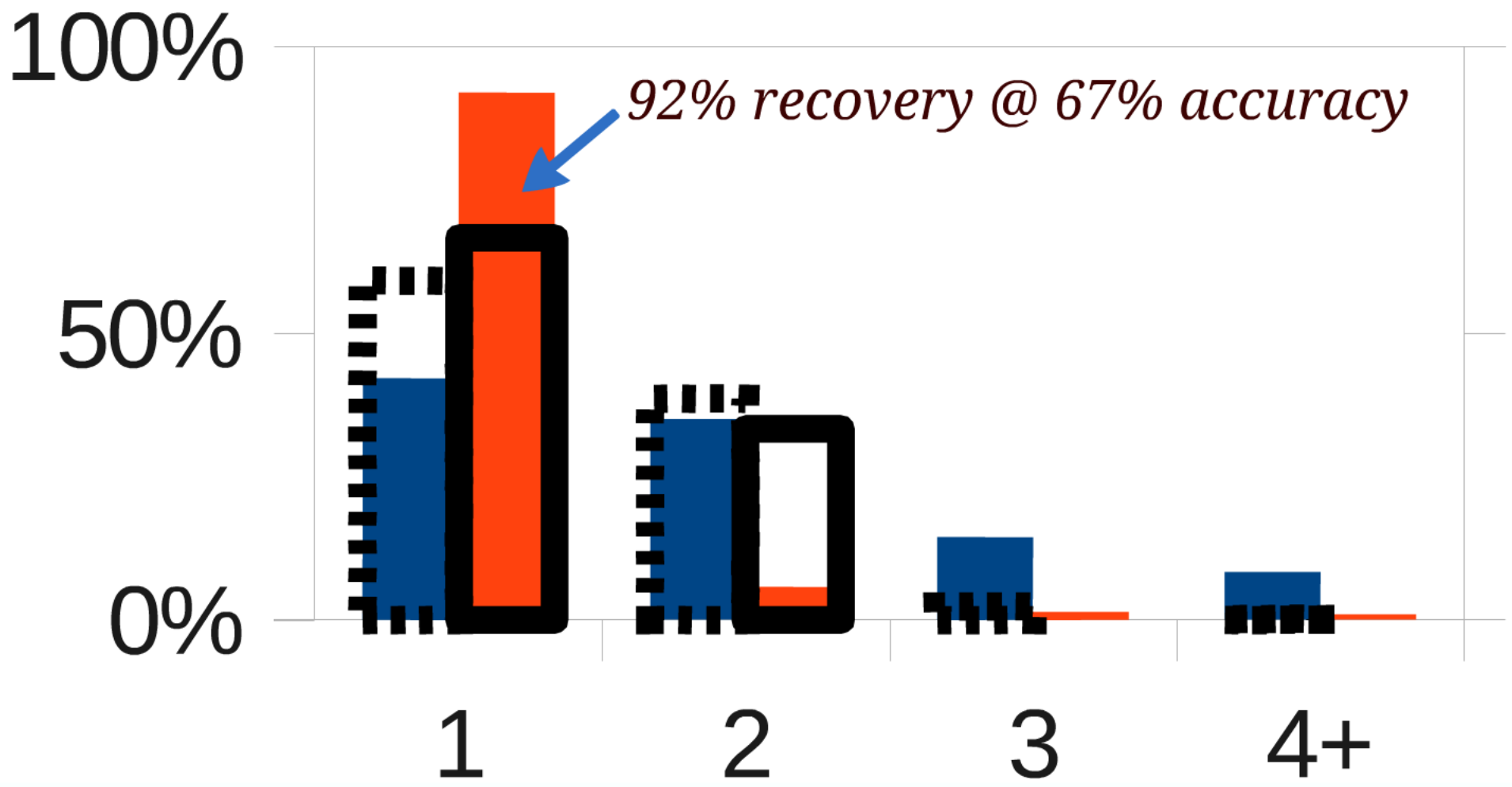
[C] ED/EC: bline,dense,pt



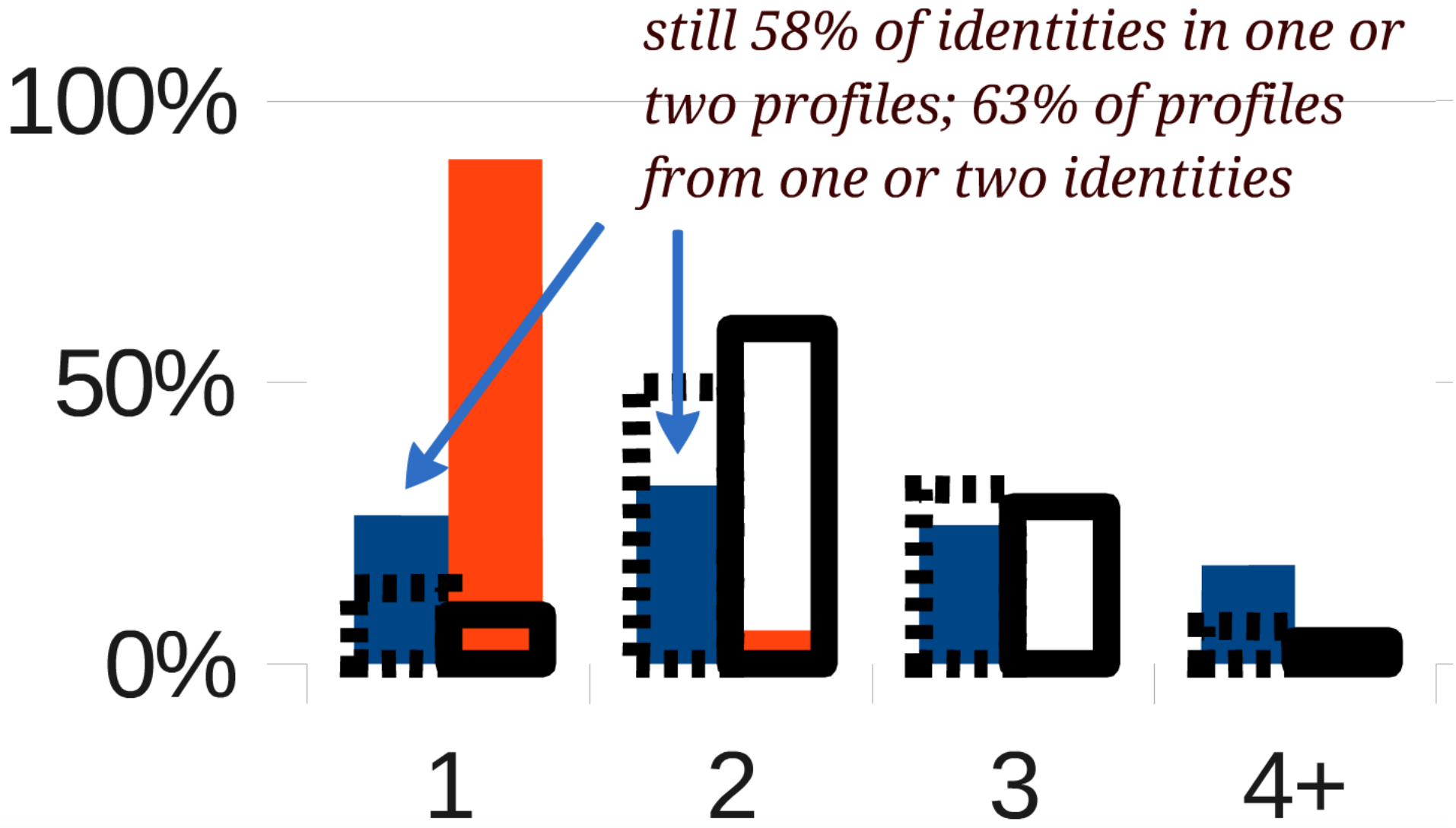
[I] ED/EC: low-o,dense,np



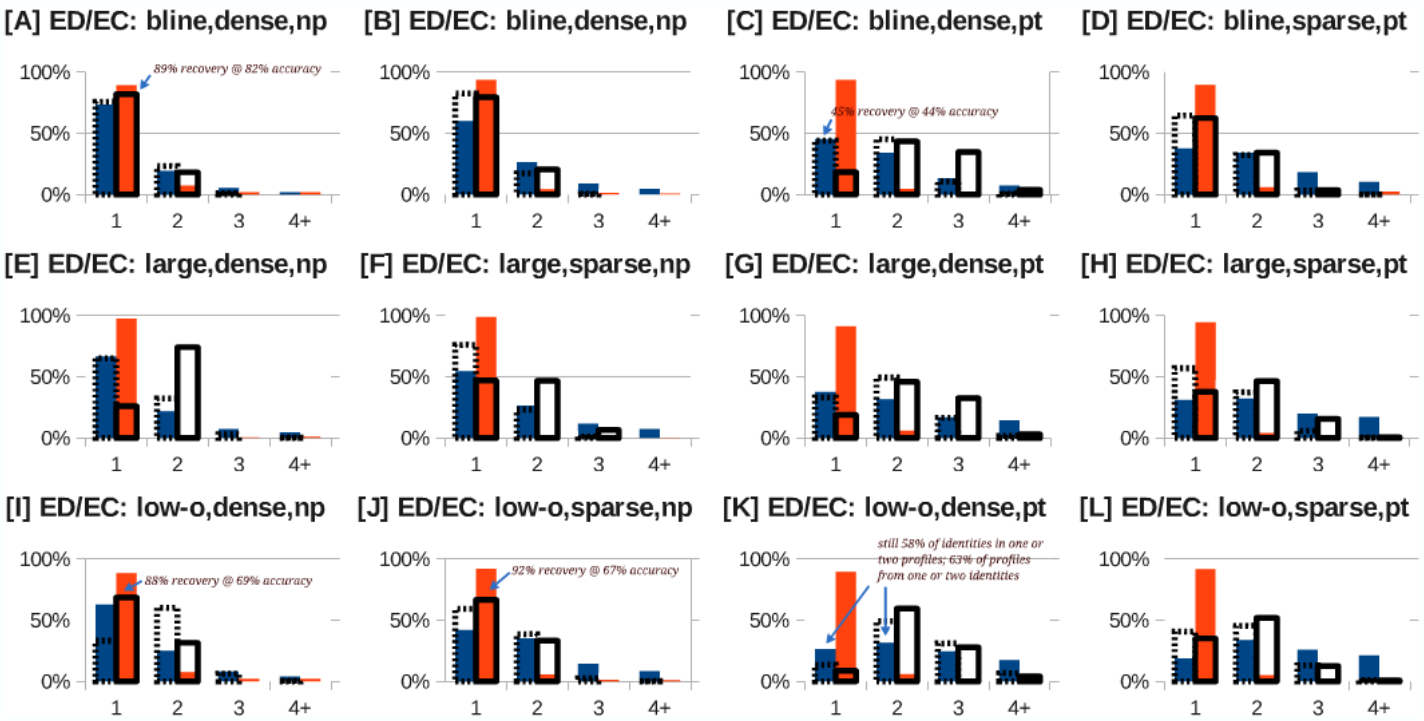
[J] ED/EC: low-o,sparse,np



[K] ED/EC: low-o,dense,pt



experimental numbers

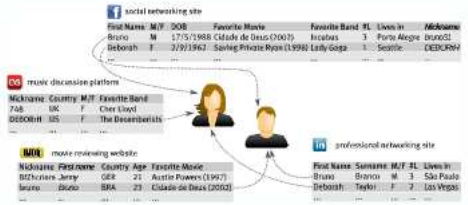


experimental numbers

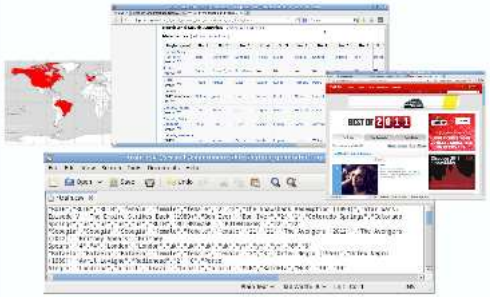
Table 1. Average precision ("p")/recall ("r")/f-measure ("f") results of our experiments after linking phase ("PW"); community detection ("CD"); and threshold transitive closure ("TC"). Boldfaced f-measures indicates significantly best result(s). "Pt" means perturbation.

	Dense, no pt			**Sparse, no pt**			**Dense, pt**			**Sparse, pt**		
	p	r	f	p	r	f	p	r	f	p	r	f
Baseline PW	0.37	0.83	0.51	0.38	0.83	0.52	0.24	0.68	0.36	0.25	0.68	0.37
CD	0.75	0.84	**0.79**	0.50	0.82	0.62	0.48	0.58	**0.52**	0.35	0.60	**0.44**
TC	0.49	0.91	0.63	0.62	0.83	**0.70**	0.57	0.30	0.39	0.28	0.60	0.38
Large PW	0.23	0.83	0.36	0.24	0.83	0.37	0.14	0.68	0.23	0.14	0.68	0.23
CD	0.64	0.76	**0.69**	0.42	0.74	0.53	0.35	0.48	**0.40**	0.26	0.51	**0.34**
TC	0.52	0.50	0.51	0.91	0.42	**0.57**	0.20	0.32	0.24	0.41	0.25	0.31
Low Overlap PW	0.26	0.67	0.38	0.26	0.67	0.37	0.14	0.43	0.21	0.14	0.44	0.21
CD	0.50	0.44	0.47	0.32	0.51	0.39	0.20	0.23	0.22	0.15	0.27	0.20
TC	0.45	0.85	**0.58**	0.58	0.71	**0.64**	0.34	0.22	**0.26**	0.38	0.22	**0.28**

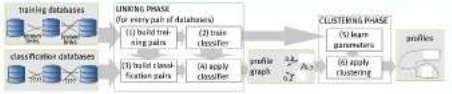
multiple databases



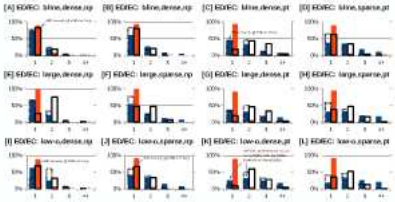
common attributes & values



experimental numbers



experimental numbers



Conclusion:
non-identifying data can often be linked & clustering helps

Future work:
overcoming our limitations:
approaches & experiments & insight

experimental numbers

Table 1. Average precision ("p")/recall ("r")/f-measure ("f") results of our experiments after linking phase ("PW"); community detection ("CD"); and threshold transitive closure ("TC"). Boldfaced f-measures indicates significantly best result(s). "Pt" means perturbation.

		Dense, no pt			Sparse, no pt			Dense, pt			Sparse, pt		
		p	r	f	p	r	f	p	r	f	p	r	f
Baseline	PW	0.37	0.83	0.51	0.38	0.83	0.52	0.24	0.68	0.36	0.25	0.68	0.37
	CD	0.75	0.84	**0.79**	0.50	0.82	0.62	0.48	0.58	**0.52**	0.35	0.60	**0.44**
	TC	0.49	0.91	0.63	0.62	0.83	**0.70**	0.57	0.30	0.39	0.28	0.60	0.38
Large	PW	0.23	0.83	0.36	0.24	0.83	0.37	0.14	0.68	0.23	0.14	0.68	0.23
	CD	0.64	0.76	**0.69**	0.42	0.74	0.53	0.35	0.48	**0.40**	0.26	0.51	**0.34**
	TC	0.52	0.50	0.51	0.91	0.42	**0.57**	0.20	0.32	0.24	0.41	0.25	0.31
Low Overlap	PW	0.26	0.67	0.38	0.26	0.67	0.37	0.14	0.43	0.21	0.14	0.44	0.21
	CD	0.50	0.44	0.47	0.32	0.51	0.39	0.20	0.23	0.22	0.15	0.27	0.20
	TC	0.45	0.85	**0.58**	0.58	0.71	**0.64**	0.34	0.22	**0.26**	0.38	0.22	**0.28**